Machine Learning Methods for Small Data Challenges in Molecular Science.

Year: 2023 | Citations: 279 | Authors: Bozheng Dou, Zailiang Zhu, E. Merkurjev, Lu Ke, Long Chen

---

Abstract

Small data are often used in scientific and engineering research due to the presence of various constraints, such as time, cost, ethics, privacy, security, and technical limitations in data acquisition. However, big data have been the focus for the past decade, small data and their challenges have received little attention, even though they are technically more severe in machine learning (ML) and deep learning (DL) studies. Overall, the small data challenge is often compounded by issues, such as data diversity, imputation, noise, imbalance, and high-dimensionality. Fortunately, the current big data era is characterized by technological breakthroughs in ML, DL, and artificial intelligence (AI), which enable data-driven scientific discovery, and many advanced ML and DL technologies developed for big data have inadvertently provided solutions for small data problems. As a result, significant progress has been made in ML and DL for small data challenges in the past decade. In this review, we summarize and analyze several emerging potential solutions to small data challenges in molecular science, including chemical and biological sciences. We review both basic machine learning algorithms, such as linear regression, logistic regression (LR), k-nearest neighbor (KNN), support vector machine (SVM), kernel learning (KL), random forest (RF), and gradient boosting trees (GBT), and more advanced techniques, including artificial neural network (ANN), convolutional neural network (CNN), U-Net, graph neural network (GNN), Generative Adversarial Network (GAN), long short-term memory (LSTM), autoencoder, transformer, transfer learning, active learning, graph-based semi-supervised learning, combining deep learning with traditional machine learning, and physical model-based data augmentation. We also briefly discuss the latest advances in these methods. Finally, we conclude the survey with a discussion of promising trends in small data challenges in molecular science.